Standardisierung von der Heterogenität her denken - zum Entwicklungsstand bilateraler Transferkomponenten für digitale Fachbibliotheken

Krause, Jürgen

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Krause, J. (2003). *Standardisierung von der Heterogenität her denken - zum Entwicklungsstand bilateraler Transferkomponenten für digitale Fachbibliotheken.* (IZ-Arbeitsbericht, 28). Bonn: Informationszentrum Sozialwissenschaften. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-50750-9

GESIS

IZ InformationsZentrum Sozialwissenschaften

IZ-Arbeitsbericht Nr. 28

Standardisierung von der Heterogenität her denken - Zum Entwicklungsstand bilateraler Transfer-komponenten für digitale Fachbibliotheken

Jürgen Krause

Juli 2003

InformationsZentrum
Sozialwissenschaften

Lennéstraße 30
D-53113 Bonn
Tel.: 0228/2281-0
Fax.: 0228/2281-120
email: iz@bonn.iz-soz.de
Internet: http://www.gesis.org/IZ/index.htm

ISSN: 1431-6943

Herausgeber: Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI)

Druck u. Vertrieb: Informationszentrum Sozialwissenschaften, Bonn
Printed in Germany

Inhalt

1 Einleitung 4

2 Heterogenitätsbehandlung zwischen Benutzer und Dokumenterschließung ist nicht neu 10

3 Bilaterale Heterogenitätsbehandlung als Basis inhomogener Dokumentensammlungen 12

4 Einzelverfahren und erreichter Stand 14

4.1 Crosskonkordanzen und qualitative Heterogenität 14

4.2 Statistische Transferverfahren 17

4.3 Statistisch ermittelte Crosskonkordanzen 19

4.4 Zusammenspiel der Einzelverfahren 20

5 Transfer-Architektur und Einbettung in dezentrale und zentralistische Ansätze 22

6 Wie konzeptualisiert der Benutzer seine Anfrage an einen dezentralen, heterogenen Informationsraum? 24

7 Transfermodule als erster Baustein eines Schalenmodells der Informationsverarbeitung 28

8 Literatur 30

1 Einleitung

Bei der Diskussion um Normierungs- und Standardisierungsbemühungen in den Bibliotheken, Archiven, den Fachdatenbanken, Texten auf WWW-Servern wissenschaftlicher Institute und anderen wissenschaftlichen Informationsservicestellen zeichnet sich immer deutlicher ein tief greifender Wandel der Rahmenbedingungen ab. Der technologische, wirtschaftliche, politische und gesellschaftliche Wandel der letzten Jahre erzeugt Randbedingungen, zu denen die in den letzten Jahrzehnten als gültig angesehenen Lösungskonzepte der Produktion, Distribution und Nutzung von wissenschaftlicher Information in Widerspruch geraten. Die Irritationen und möglichen Konsequenzen lassen sich besonders deutlich am integrierenden Konzept der digitalen Fachbibliotheken festmachen. Digitale (Fach)bibliothek - wie der Term bei Global-Info, den DFG-Sonderprogrammen und in der Literatur verwendet wird - bedeutet: Wissenschaftler sollen von ihrem Computer aus einen optimalen Zugang zu den weltweit vorhandenen elektronischen und multimedialen Volltext-, Literaturhinweis-, Fakten- und WWW-Informationen haben, einschließlich der dort vorhandenen Lehrmaterialien, Spezialverzeichnisse zu Experten etc.

Auf technischer Seite setzt dies im Netz zugängliche verteilte Datenbanken voraus, auf konzeptueller Seite die Integration verschiedener Informationsgehalte und -strukturen. Traditionell wird versucht, konzeptuelle Integration durch Standardisierung und Normierung sicherzustellen. Wissenschaftler, Bibliotheken, Verleger und die Anbieter von Fachdatenbanken müssten sich z. B. auf Dublin Core-Metadaten und eine einheitliche Klassifikation wie die DDC einigen. Damit würden homogene Datenräume geschaffen, deren Konsistenz qualitativ hochwertige Recherchen erlaubt.

Im Sinne des DIN-Papiers „Strategie für die Standardisierung der Informations- und Kommunikationstechnik“ (DIN-SICT 2002) sind Standards „jedwede in einem Konsensprozess entstandenen Spezifikationen, wobei es hinsichtlich des Konsensrahmens beträchtliche Abstufungen geben kann.“

Bei den Normen kommt die Konsensbildung „in einem öffentlichen Einspruchsverfahren“ hinzu. Nach DIN 820 (1994) ist Normung „die planmäßige, durch die interessierten Kreise gemeinschaftlich durchgeführte Vereinheitlichung von materiellen und immateriellen Gegenständen zum Nutzen der Allgemeinheit. Sie darf nicht zu einem wirtschaftlichen Sondernutzen Einzelner führen. Sie fördert die Rationalisierung und Qualitätssicherung in Wirtschaft, Technik, Wissenschaft und Verwaltung. Sie dient der Sicherheit von

Menschen und Sachen sowie der Qualitätsverbesserung in allen Lebensbereichen. Sie dient außerdem einer sinnvollen Ordnung und der Information auf dem jeweiligen Normungsgebiet."

Es gibt klare Anzeichen dafür, dass die traditionellen Verfahren der Standardisierung und Normierung an ihre Grenzen stoßen. Einerseits erscheinen sie unverzichtbar und haben in Teilbereichen deutlich die Qualität der wissenschaftlichen Informationssuche gesteigert. Andererseits sind sie im Rahmen globaler Anbieterstrukturen von Informationen nur noch partiell durchsetzbar, bei steigenden Kosten. Besonders im Teilbereich der Inhaltserschließung wird deutlich, dass für digitale Fachbibliotheken - bei allen notwendigen Bemühungen - nicht von der Durchsetzbarkeit einheitlicher Standards der Inhaltsbeschreibung ausgegangen werden kann. Es wird immer Anbieter geben, die sich den vorgegebenen Standards der Mehrheit nicht unterordnen, auf deren Daten der Benutzer im Rahmen einer integrierten Recherche jedoch nicht verzichten möchte.

Da digitale Fachbibliotheken hybride Bibliotheken aus elektronischen und gedruckten Texten und Informationen sind, treffen in ihnen die Regeln und Standards aufeinander, die für diese beiden Welten jeweils Gültigkeit haben. Sie werden im Detail in Krause/Niggemann/Schwänzl 2003 besprochen. Diese Bestandsaufnahme der Normierungs- und Standardisierungsbemühungen in den einzelnen Teilbereichen einer digitalen Fachbibliothek zeigt folgende Tendenzen:

- Alle Bemühungen der inhaltlichen Erschließung der Informationsbestände folgen der klassischen Normierungs- und Standardisierungsphilosophie, der auch die theoretischen Grundlagen der Inhaltserschließung in den Informations- und Bibliothekswissenschaften verhaftet ist. Nach einem normierten, intellektuell kontrollierten Verfahren, das eine Zentralstelle entwickelt und durchsetzt, erfolgt eine einheitliche Erfassung der Dokumente. In diesem Denken kommt der Datenkonsistenz die höchste Priorität zu, wodurch der Benutzer (idealiter) immer einem homogenisierten Datenbestand gegenüber steht. Darauf war die gesamte IuD-Methodik, einschließlich der Verwaltungsstruktur der Zentren, ausgerichtet. Möglichst weitgehende Regulation soll die für Benutzerfragen als notwendig angesehene Konsistenz erreichen. Dieses Modell der Inhaltserschließung, das z. B. auch allen Bibliothekserschließungen zu Grunde liegt, erwies sich in Teilbereichen wie bei den OPACs und Fachdatenbanken durchaus als ein gangbarer Weg, der sich in den letzten zwanzig Jahren bewährte. Was sich jedoch geändert hat, sind die Rahmenbedingungen. Der technologische, wirtschaftliche, politische und gesellschaftliche Wandel der letzten Jahre brachte Strömungen

und Meinungen hervor, die zu diesem Modell in einigen Punkten in Widerspruch gerieten (siehe auch Krause 1999).

- Trotz aller Plausibilität der damit verbundenen Vorteile ist es auch in den Bibliotheken nur in Teilbereichen gelungen, den Standardisierungsanspruch wirklich durchzusetzen. Spätestens seit dem Aufkommen der Fachdatenbanken, die sich die Erschließung von Zeitschriftenartikeln zur zentralen Aufgabe machten, müssen Benutzer mit unterschiedlichen Erschließungskonzepten arbeiten. Bei den für digitale Fachbibliotheken angestrebten Teilkomponenten sind heute in allen Fächern eine Vielzahl von Sacherschließungssystemen vertreten. Das Spektrum reicht z. B. bei den Sozialwissenschaften von allgemeinen Regelwerken, wie der Schlagwortnormdatei (SWD - ist die von den deutschen wissenschaftlichen Universalbibliotheken kooperativ aufgebaute Schlagwortnormdatei auf der Basis des Regelwerks RSWK „Regeln für den Schlagwortkatalog“, siehe Krause/Niggemann/Schwänzl 2003), über fachspezifische, wie dem Thesaurus Sozialwissenschaften des IZ), bis zu freien Schlagworten. Gleiches gilt für die Klassifikationen, die mit verschiedenen Ausprägungen der Basisklassifikation (BK) als Allgemeinklassifikation, der Klassifikation Sozialwissenschaften (IZ-Klassifikation) als Fachklassifikation und den weniger spezifischen Aufstellsystematiken einzelner Bibliotheken vertreten sind.

- Die WWW-Institutsserver der Universitäten, auf denen Wissenschaftler ihre Forschungsergebnisse in verschiedenster Form zur Verfügung stellen, haben diese Grundsituation verschärft, obwohl auch hier internationale Standardisierung zur Konsistenzerhaltung die Leitidee war. Durch neue Metadaten-Ansätze wie die Dublin Core Initiative (DC) soll die im WWW verloren gegangene Konsistenz aus der Welt der Bibliotheken und Fachdatenbanken teilweise neu etabliert werden. Im Gegensatz zu den Bibliotheken und Fachdatenbanken sind die Ziele der Etablierung von WWW-Metadaten bescheidener: Hier sind sie Übereinkünfte, bestimmte Merkmale eines Dokumentbestandes in einer verabredeten Form bei den eigenen Daten auszuweisen, wie verschieden sie in Bezug auf andere Merkmale auch immer sein mögen. Aber auch in diesem bescheideneren Kontext geht man davon aus, dass Standardisierungsbemühungen nur partiell erfolgreich sein können. Auch im günstigsten Fall werden sich nicht alle WWW-Anbieter z. B. auf eine Form der Inhaltserschließung einigen.

- Digitale Fachbibliotheken wollen nicht nur textuelle Informationen erschließen, sondern idealiter gleichzeitig Fakten (z. B. die Zeitreihen von Umfragen) oder auch Lehrmaterialien und andere Medialitäten wie Bild, Ton, Animation oder Videosequenzen nachweisen. Auch hier wird der In-

halt in verschiedenen Formen ausgedrückt. Ein gemalter Baum kann die gleiche Semantik haben wie ein fotografierter oder wie das gedruckte oder auch gesprochene Wort „Baum“. Das Datum bedient sich jeweils eines anderen Notationssystems. Bei Text versus Bild wird das sofort sichtbar, bei inhaltlich verschieden erschlossenen Dokumenten ist der gleiche Sachverhalt versteckter, weil auf der Oberfläche im Einzelfall die gleiche Buchstabensequenz verwendet werden kann. Dennoch handelt es sich um ein konzeptuelles Kontinuum, nicht um zwei getrennte Phänomene.

- Die Konsequenzen aus den obigen Beobachtungen sind für die digitalen Fachbibliotheken gravierend. Gerade durch den an mehreren Stellen wie dem GBV (Gemeinsamer Bibliotheksverbund) oder KOBV (Kooperative Bibliotheksverbund Berlin – Brandenburg) erfolgten technologischen Zusammenschluss zeigen sich die verschiedensten Konsistenzbrüche zwischen den integrierten Teilbeständen.

- Relevante, qualitätskontrollierte Daten stehen neben irrelevanten und eventuell nachweislich falschen. Nur noch in abgegrenzten Teilbereichen sorgen Gutachtersysteme für eine Trennung von Ballast und potentiell erwünschter Information.

- Ein Deskriptor A kann in einem solchem System die unterschiedlichsten Bedeutungen annehmen. Auch im engen Bereich der Fachinformation kann ein Deskriptor A, der aus einem hochrelevanten Dokumentenbestand mit viel Aufwand intellektuell und qualitativ hochwertig ermittelt wurde, nicht mit dem Term A gleichgesetzt werden, den eine automatische Indexierung aus einem Randgebiet liefert.

Kaum jemand hängt heute noch der Vorstellung nach, der Dokumentenraum ließe sich durch globale Standardisierungsabsprachen über alle Teilbereiche hinweg homogenisieren, sich organisatorisch wieder auf einige wenige Mitspieler reduzieren oder über ein hierarchisch organisiertes Modell der Kooperation gestalten. Ganz im Gegenteil, die heutigen Vorstellungen gehen von einer noch stärkeren Dezentralisierung bei der Dokumenterstellung, -erschließung und -verteilung aus, wodurch die „anarchischen Tendenzen“ weiter zunehmen. Der Benutzer wird trotz solcher Probleme auf alle Dokumentenbestände zugreifen wollen, gleich nach welchen Verfahren sie erschlossen oder in welchem System sie angeboten werden. Er hält auch in der Welt dezentralisierter, inhomogener Dokumentenbestände die Forderung an die Systementwickler aufrecht, dafür zu sorgen, dass er möglichst nur die relevanten Dokumente und möglichst alle relevanten nachgewiesen bekommt.

Deshalb muss der verbleibenden und unvermeidlichen Heterogenität durch verschiedene Strategien begegnet werden. Neue Problemlösungen und Weiterentwicklungen sind in beiden Bereichen nötig:

- den Metadaten
- und den Methoden des Umgangs mit der verbleibenden Heterogenität. Zwischen den einzelnen Datentypen (z. B. Literaturdatenbanken und Internetquellen) sind aufeinander abgestimmte Transfermodule zu spezifizieren, die drei Methodenklassen zuzuordnen sind:
  - Crosskonkordanzen und -klassifikationen als konzeptuell einfachste Form des Transfers, der aber nur bei auf der Basis des Wortschatzes generalisierbaren Relationen wirksam wird
  - sowie quantitativ-statistische und
  - deduktive Ansätze.

Beide Anforderungen hängen eng zusammen. Durch die Fortentwicklung im Bereich der Metadaten soll einerseits die verloren gegangene Konsistenz partiell hergestellt werden. Andererseits sollen mit Verfahren zur Behandlung von Heterogenität Dokumente unterschiedlichen Niveaus der Datenrelevanz und Inhaltserschließung aufeinander bezogen werden. Diese Vorgehensweise lässt sich durch die folgende Prämisse umschreiben: Standardisierung ist von der verbleibenden Heterogenität her zu denken. Erst im gemeinsamen Zusammenwirken von intellektuellen und automatischen Verfahren zur Heterogenitätsbehandlung und Standardisierung ergibt sich eine Lösungsstrategie, die den heutigen technischen, politischen und gesellschaftlichen Rahmenbedingungen gerecht wird.

Die Forschungs- und Entwicklungsabteilung des Informationszentrums Sozialwissenschaften (IZ Bonn) hat in den letzten Jahren die theoretischen Grundlagen dieser Sichtweise auf die Standardisierungsprobleme und ihre praktische Umsetzung in verschiedenen Projekten zusammen mit unterschiedlichen Partnern bearbeitet:

ELVIRA: **„El**ektronisches **V**erbands**i**nformations-, **R**echerche- und –**A**nalysesystem“; gefördert durch das Bundesministerium für Wirtschaft (BMWi)

Partner: Verband Deutscher Maschinen- und Anlagenbau e. V. (VDMA); Zentralverband der Elektrotechnik- und Elektronikindustrie e. V. (ZVEI); Hauptverband der Deutschen Bauindustrie e. V. (HVB); Deutsches Institut für Wirtschaftsforschung (DIW), Berlin; ifo-Institut für Wirtschaftsforschung,

München (siehe Stempfhuber/Hellweg/Schaefer 2002, Stempfhuber 2003, URL: http://www.gesis.org/Forschung/Informationstechnologie/ELVIRA.htm)

CARMEN: „**C**ontent **A**nalysis, **R**etrieval and **M**etadata: **E**ffective **N**etworking“; gefördert durch das BMBF

Partner: Universität Osnabrück, Fachbereich Mathematik/Informatik; Universität Regensburg, Universitätsbibliothek; Die Deutsche Bibliothek (DDB), Frankfurt/Main; Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Frankfurt/Main (siehe Krause/Plümer/Schwänzl 2000, Strötgen 2002, URL: http://www.gesis.org/Forschung/Informationstechnologie/CARMEN-AP11.htm)

ViBSoz: Digitale Fachbibliothek Sozialwissenschaften, DFG-Förderung

Partner: TU Darmstadt, Institut für Soziologie; Friedrich-Ebert-Stiftung (FES); Universitäts- und Stadtbibliothek (USB) Köln; Wissenschaftszentrum für Sozialforschung Berlin (WZB) (siehe Müller/Marx (2001),
URL: http://www.gesis.org/Forschung/Informationstechnologie/ViBSoz.htm)

ETB: „The European Schools Treasury Browser“, gefördert durch die EU

Partner: Humboldt-Universität Berlin, Abt. Pädagogik und Informatik; Biblioteca di Documentazione Pedagogica Firenze (BDP); Universidad National de Educación a Distanzia Madrid (UNED); Lunds Universiteit – NetLab Lund, Schweden (siehe Kluck 2001, Kluck/Strötgen 2002,
URL: http://www.gesis.org/Forschung/Informationstechnologie/ETB.htm,
URL: www.eun.org/etb/)

infoconnex: Informationsverbund Pädagogik – Sozialwissenschaften – Psychologie. Aufbau eines Volltextservers Zeitschriften für Pädagogik, Psychologie, Sozialwissenschaften; BMBF- und DFG-Förderung

Partner: Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Frankfurt/Main; Zentrum für Psychologische Information und Dokumentation (ZPID), Trier; Universitätsbibliothek Erlangen-Nürnberg; Universitäts- und Stadtbibliothek Köln (USB Köln); Saarländische Universitäts- und Landesbibliothek (SULB), Saarbrücken
URL: www.infoconnex.de, www.iwi-iuk.org/iuk2003/program

In den Projekten entwickelten sich sowohl die Modellvorstellungen als auch deren praktische Umsetzung kontinuierlich weiter (siehe als Einstieg die IZ-Arbeitsberichte 6, 17, 19, 21 – 24 unter
http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/; als Überblick über die Forschungsprogramme des BMBF und der DFG siehe Schö-

ning-Walter 2003). Der im Folgenden dargestellte Stand spiegelt das Ergebnis der Entwicklung der letzten sechs Jahre wider und fasst die realisierten Teilkomponenten solch einer Modellsicht zusammen. Das bisher Erreichte bietet schon heute erste Ansätze, die das konzeptuelle Problem unterschiedlichster Inhaltserschließungen im Kontext der heutigen digitalen Fachbibliotheken, Informationsverbünde und Fachportale zumindest partiell lösen.

2 Heterogenitätsbehandlung zwischen Benutzer und Dokumenterschließung ist nicht neu

Jedem Bibliothekar und jedem, der sich mit Information Retrieval befasst, war schon immer klar, dass zwischen den semantischen Termen, die in der Datenbank ein Dokument charakterisieren und dem Term, den der Benutzer anwendet, nicht immer eine 1:1-Relation besteht. Benutzer gebrauchen für die gleiche Semantik andere Terme als die Bearbeiter bei der Inhaltserschließung. Auch wenn der Benutzer sich des Standards der Inhaltserschließung über einen Fachthesaurus bewusst ist, wird er nicht bei jeder Anfrage, die ihm durch den Kopf geht, den entsprechenden Thesaurusbegriff nachschlagen oder Abweichungen zu seiner Alltagssprache im Kopf haben. Unklar ist ihm oft auch, ob der Thesaurus prä- oder postkoordiniert ist, ob somit ein Kompositum wie *Jugendarbeitslosigkeit* verwendet werden kann wie bei der SWD oder in *Jugend* und *Arbeitslosigkeit* zu zerlegen ist wie bei SOLIS (**So**zialwissenschaftliches **Li**teraturinformations**s**ystem, http://www.gesis.org/Information/SOLIS/index.htm).

Eingesetzte Gegenstrategien zu dieser jeder traditionellen Standardisierungslösung inhärenten „Heterogenität“ zwischen Erschließungs- und Benutzerebene waren:

- Der Thesaurus wird durch Oberbegriffs-/Unterbegriffsrelation, durch Synonyme, Ähnlichkeitsbegriffe usw. angereichert.

- Linguistische Komponenten wie Morphologie, Derivation oder semantische Begriffsnetze tragen dem Umstand Rechnung, dass Terme sich bei ihrer Sprachverwendung regelgeleitet verändern und die Kenntnis dieser Regeln diese Art der Heterogenität auch automatisch auflösen können.

- Verfahren des sog. intelligenten Information Retrieval (siehe Belkin 1996, Ingwersen 1996) gehen partiell durch komplexere qualitativ-deduktive

Regelwerke über rein linguistische Zusammenhänge hinaus und integrieren semantisch-pragmatisches Hintergrundwissen (Expertensysteme).

All diese traditionellen Vorgehensweisen können als Termerweiterungsverfahren (einschließlich Termveränderung) in die klassischen Bibliotheks- und Information Retrieval-Systeme integriert werden. Komplexere Beispiele enthält z. B. OSIRIS (Ronthaler/Zillmann 1998, Zillmann 1997), linguistische Funktionen u. a. IDX. Letzteres System wurde in den MILOS-Projekten für die maschinelle Indexierung von Bibliotheksbeständen eingesetzt (http://www.uni-duesseldorf.de/ulb/mil_kurz.htm). IDX ist ein wörterbuchbasiertes Verfahren für die Sprachen Deutsch, Englisch und Französisch, das zu den im Text vorkommenden Wortformen die Grundformen ermittelt und Mehrworterkennung sowie Kompositazerlegung unterstützt.

Generell lassen sich auch solche Techniken als Vagheitsproblem zwischen Benutzeranfrage und Datenbank modellieren. Diese Sichtweise stammt aus der Welt der statistisch-quantitativen IR-Verfahren (meist probabilistisches und Vektorraummodell), wobei die Vagheit zwischen Benutzeranfrage und Dokumentenbestand modelliert wird. Eher am Rande stehen Versuche, statt statistischer Verfahren neuronale Netze einzusetzen (s. Mandl 2001 als Überblick).

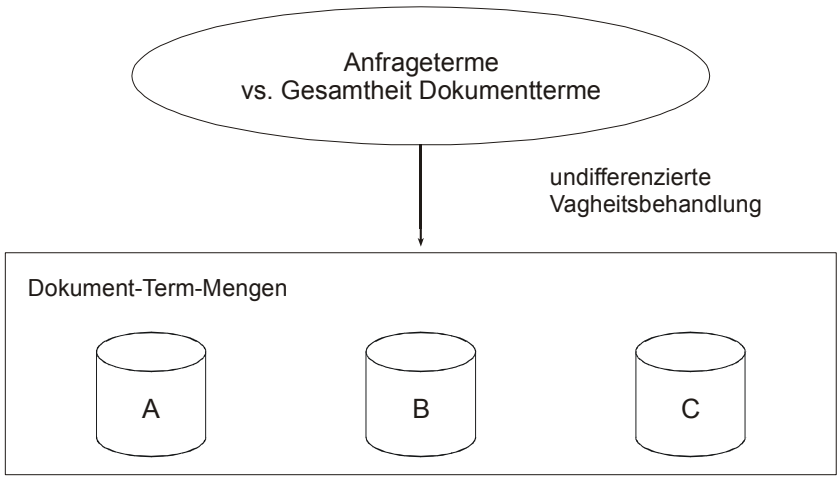

Abb. 1: Traditionelle undifferenzierte Vagheitsbehandlung

Die traditionelle Form der Vagheitsbehandlung im Information Retrieval (IR) bezieht sich somit auf den Vergleich der Anfrageterme mit denen der Inhaltserschließung der Einzeldokumente, wobei die Dokumentenebene als einheitlich zu modellierende Menge verstanden wird. Am deutlichsten zeigt sich

dies, wenn die Dokumente zusätzlich zur intellektuellen Inhaltserschließung (Schlagworte, Deskriptoren) automatisch indexiert werden (Stichworte, Freitextanalyse). Auch dann folgt die Modellierung der Anfrageseite im Prinzip dieser Homogenitätsforderung. Der Benutzer kann entweder mithilfe der Deskriptoren suchen und seine Suchstrategie an dem kontrollierten Vokabular und der intellektuellen Erschließung ausrichten oder aber mithilfe der Stichworte mit einer anderen Suchstrategie (oder beide Felder bedienen, was dann bei der Anfrageformulierung zwei getrennten Umsetzungen des Informationsbedürfnisses entspricht). Es gibt keinen systemseitigen Überstieg von der einen Gruppe in die andere.

3 Bilaterale Heterogenitätsbehandlung als Basis inhomogener Dokumentensammlungen

Solange die Dokumente aus einer homogenen Datenbank stammen, spricht nichts gegen die traditionelle Vorgehensweise. Dies ändert sich erst bei der integrierten Recherche über heterogene Dokumentenbestände, wenn wir z. B. bei digitalen Bibliotheken versuchen, eine sozialwissenschaftliche Literaturdatenbank wie SOLIS mit ihrem eigenen kontrollierten Vokabular (Thesaurus Sozialwissenschaften des IZ und IZ-Klassifikation) mit einem Bibliothekskatalog - z.B. im Projekt ViBSoz mit dem der Universitätsbibliothek Köln - zu verbinden, dessen Dokumente intellektuell nach der SWD erschlossen werden. Vergleicht man zwei solche Thesauri oder Klassifikationen, wird deutlich, dass die Vagheit hier bereits auf der semantischen Beschreibungsebene der Dokumentenbestände entsteht, nicht erst bei der Benutzeranfrage. Z. B. ist die Bedeutung eines Deskriptors X der Bibliotheksklassifikation eine andere als die Bedeutung des gleichen Terms in einer anderen Klassifikation oder auch in einem Thesaurus wie dem Thesaurus Sozialwissenschaften des IZ zur sozialwissenschaftlichen Literaturdatenbank SOLIS, wenn auch ein - eben „vager“ - Zusammenhang besteht. Bei den heutigen Verfahren lässt man diese Vagheitsrelationen undifferenziert in die Modellierung des IR-Prozesses einfließen. Man modelliert die Vagheit zwischen Benutzeranfrage und Dokumenten, ohne die Differenzen auf der Dokumentenebene zwischen jeweils zwei heterogen erschlossenen Dokumentenbeständen explizit durch Transformationsmodule gesondert zu behandeln.

Je mehr heterogene Dokumentenbestände gleichzeitig die Basis einer Recherche bilden sollen, umso schädlicher wirkt sich die Undifferenziertheit bei der

Modellierung aus. Deshalb basieren alle Transferverfahren der IZ-Projekte auf einer bilateralen Modellierung der Vagheit, deren Ergebnisse erst in weiteren Schritten kaskadierend zusammenfließen. Die These ist, dass heterogene Dokumentenbestände zuerst durch Transfermodule bilateral miteinander verbunden werden sollten (Vagheitsmodellierung auf Dokumentenebene), bevor sie in den übergeordneten Prozess der Vagheitsbehandlung zwischen Dokumenten und Anfrage (das klassische IR-Problem) eingefügt werden.

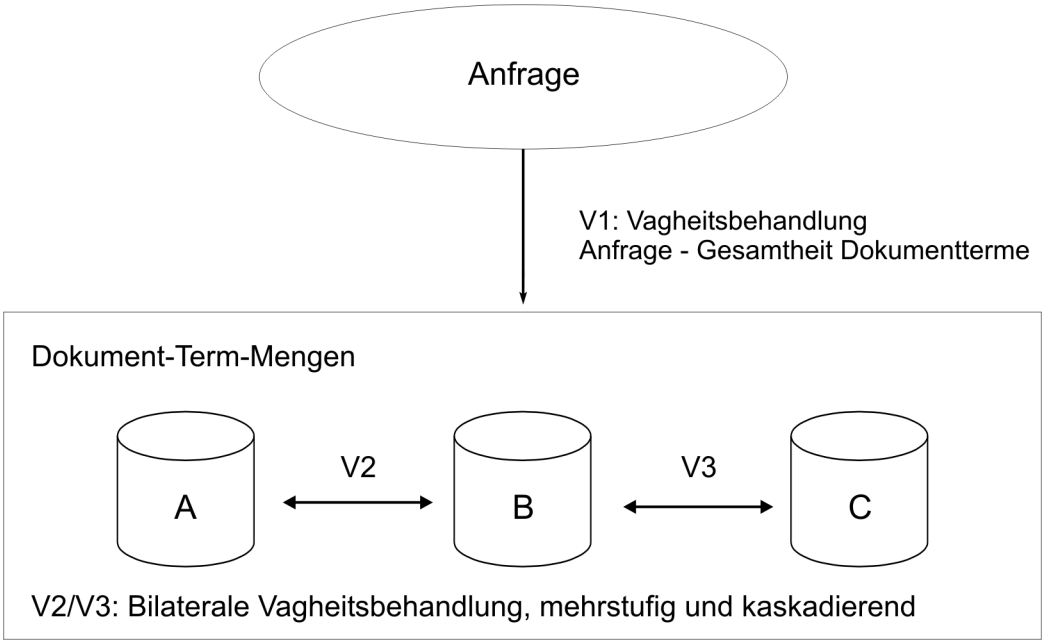

Abb. 2: Bilaterale Heterogenitätsbehandlung

Sind z. B. - wie in Abbildung 2 - drei heterogene Dokumentenbestände zu integrieren, behandeln Transfermodule zwischen A - B und B - C jeweils bilateral die Vagheit zwischen den verschiedenen Inhaltserschließungsverfahren. Die Hoffnung hinter dieser - von der traditionell im IR praktizierten Vorgehensweise deutlich abweichenden Form - ist, dass sich durch die Trennung des Vagheitsproblems in bilaterale Segmente eine größere Flexibilität und Zielgenauigkeit des Gesamtverfahrens ergibt. Verschiedene Formen der Vagheit können nahe an der verursachenden Schnittstelle (z. B. den Unterschieden zweier Thesauri) behandelt werden. Dies erscheint zum einem kognitiv plausibler und ermöglicht zum anderen die Kombination verschiedenster Methodiken zur Vagheitsbehandlung (probabilistisches Modell, neuronale Netzwerke, intellektuell erstellte Crosskonkordanzen), die gemeinsam beim Retrieval über heterogene Datenbestände wirksam werden können.

Gerade vor dem breiten Hintergrund der empirischen Hinweise der TREC- und DELOS-Studien (http://trec.nist.gov/ und http://www.sztaki.hu/delos_wg21/), dass sich die IR-Verfahren stärker in der Ergebnismenge als in der Güte der Evaluationsparameter wie Recall und Precision unterscheiden (s. Mandl 2001), scheint dieses Vorgehen vielversprechend. So kann z. B. ein probabilistisches Verfahren beim Match zwischen Benutzeranfrage und Dokumenten mit neuronalen Transformationsnetzwerken verbunden werden oder letztere auch mit IR-Verfahren auf der Basis der Booleschen Algebra. Aber auch wenn die Transformationsmodule zwischen den heterogenen Dokumentenbeständen die gleichen Ähnlichkeitsfunktionen einsetzen wie auf der IR-Ebene zwischen Benutzeranfrage und Dokumenten, werden sich die Ergebnisse zwischen traditionellem unspezifizierten Verfahren und der bilateralen Vorgehensweise deutlich unterscheiden.

4 Einzelverfahren und erreichter Stand

Als prinzipielle Verfahren zur bilateralen Bearbeitung konzeptueller Heterogenität wurden bisher Crosskonkordanzen und statistische Verfahren bis zur Einsatzreife entwickelt. Erste Versuche gab es zum Einsatz neuronaler Netze (siehe Mandl 2001 und Krause 2000). Sie bleiben hier genauso ausgeklammert wie die deduktiven Verfahren, die am Problem der automatischen Metadatenextraktion im Projekt CARMEN erprobt wurden (s. Binder et al. 2002: Abschnitt 5.2). Bei der Metadatenextraktion haben die Transfermodule einen anderen Stellenwert in der Gesamtarchitektur. Man zieht sie heran, um fehlende Informationen (z. B. Metatag „Titel“) direkt bei der Dokumentenbeschreibung zu ergänzen.

Ausgeklammert bleiben auch weitergehende Verfahren wie die in infoconnex geplanten Autorennetzwerke (s. Mutschke 2000) oder Cognitive-Mapping Ansätze (s. Mutschke/Quan-Haase 2001). Sie liefern nicht nur Relationen zwischen einzelnen Termen, sondern verknüpfen Begriffsgruppen mit zentraler Semantik zu semantischen Strukturen.

4.1 Crosskonkordanzen und qualitative Heterogenität

Crosskonkordanzen sind intellektuell erstellte Verbindungen zwischen Termen zweier Thesauri oder Klassifikationen. Sie enthalten neben reinen Synonymrelationen auch Oberbegriffs-/Unterbegriffs- und Ähnlichkeitsrelationen.

Die verschiedenen Begriffssysteme werden im Anwendungskontext analysiert und der Versuch gemacht, ihre generelle semantische Begrifflichkeit intellektuell aufeinander zu beziehen. Beispiele hierfür aus der Crosskonkordanz zwischen SWD und dem Thesaurus Sozialwissenschaften des IZ sind:

Geschlechterrolle	Geschlechtsrolle
Sowjetunion	UdSSR
Berufliche Fortbildung	Berufliche Weiterbildung
Haushalt	Privathaushalt
Führung	Personalführung
Jugendarbeitslosigkeit	Jugendlicher + Arbeitslosigkeit

Abb. 3: Beispiele intellektuell erstellter Crosskonkordanzen

Das Konzept der Crosskonkordanzen darf nicht mit dem der Metathesauri verwechselt werden. Crosskonkordanzen streben keine neue Standardisierung bestehender Begriffswelten an. Sie erfassen die partielle Verbindung zwischen bereits bestehenden Terminologiesystemen, deren Vorarbeit genutzt wird und decken damit den statisch bleibenden Teil der Transferproblematik ab. Bei der Recherche bieten solche Verzeichnisse die Möglichkeit, Terme des einen Begriffssystems in die des anderen umzusetzen, im einfachsten Fall im Sinne einer Synonymie- oder Ähnlichkeitsrelation, aber auch als 1:n- oder n:m-Relation oder als deduktive Regelbeziehung (letztere beiden Formen wurden bisher noch nicht umgesetzt).

Bei Crosskonkordanzen spielt es im Gegensatz zur Planung von Metathesauri keine Rolle, ob sich 90 % der Begriffe oder nur 20 % aufeinander abbilden lassen, weil z. B. zwischen fachverwandten Thesauri die semantischen Überschneidungen nicht zahlreicher sind (in CARMEN getestet für Mathematik – Physik) oder weil es sich bei einem Thesaurus um ein sehr generelles, alle Fächer übergreifendes Begriffssystem handelt (wie die SWD), das mit einem tief erschließenden Fachthesaurus verbunden wird. Dies spielt schon deshalb keine Rolle, weil immer daran gedacht wird, verschiedene Typen von Transfermodulen gleichzeitig einzusetzen, so dass nicht ein Modul die gesamte Transferlast tragen muss (siehe Abschnitt 5).

Ein interessanter Nebeneffekt der Erarbeitung von Crosskonkordanzen ist das Aufdecken von Ähnlichkeiten zwischen miteinander konkurrierenden Begriffssystemen. Die für Crosskonkordanzen notwendige Analyse zeigt sehr genau auf, wie hoch der Anteil gleicher Begriffssegmente ist, aller hochemotionaler Schulenbildung für die eine oder andere Lösung zum Trotz.

Intellektuell erzeugte Crosskonkordanzen für Thesauri und für Klassifikationen wurden bisher in ELVIRA, CARMEN, ViBSoz und infoconnex entwickelt. Für textuelle Informationen zur Verfügung stehen (zum Teil beschränkt auf Auszüge der Gesamtmenge):

- Thesaurus Sozialwissenschaften des IZ ↔ SWD (CARMEN, infoconnex)
- DIPF-Thesaurus Bildung ↔ SWD (CARMEN, infoconnex)
- DIPF-Thesaurus Bildung → ZPID PSYNDEX (infoconnex)
- Thesaurus Sozialwissenschaften des IZ → Thesaurus Bildung (infoconnex)
- ZPID-Psychologie PSYNDEX ↔ Thesaurus Sozialwissenschaften des IZ (infoconnex)
- Basisklassifikation ↔ Klassifikation Sozialwissenschaften des IZ (CARMEN)
- Regensburger Verbundklassifikation (RVK) ↔ Klassifikation Sozialwissenschaften des IZ (CARMEN)
- MSC Mathematik ↔ PACS Physik (CARMEN)
- RVK ↔ PACS (CARMEN)
- RVK ↔ MSC (CARMEN)
- RVK ↔ BK (Teilgebiete Physik und Mathematik, CARMEN)
- ETB ↔ EUN (ETB)
- ETB ↔ EET (ETB)

Die Text-Fakten-Integration behandelte bisher nur ELVIRA. Es wurden Nomenklaturen und die Länderliste des Statistischen Bundesamtes und der Verbände verbunden (z. B. im Elektrobereich: Außenhandel EU-Länder - alte Kennziffern bis 1996 ↔ Außenhandel EU-Länder - neue Kennziffern ab 1996).

In die derzeit freigegebene ViBSoz-Recherche integriert sind Crosskonkordanzen zwischen der SWD und dem Thesaurus Sozialwissenschaften des IZ

sowie zwischen der Basisklassifikation und der Klassifikation Sozialwissenschaften des IZ.

Der Volltextzugang von infoconnex wird Crosskonkordanzen enthalten, die ausgehend von den drei Partnerthesauri des IZ Sozialwissenschaften, DIPF und ZPID in Richtung SWD entwickelt wurden. Weiterhin werden in infoconnex Crosskonkordanzen zwischen diesen drei Fachthesauri implementiert.

4.2 Statistische Transferverfahren

Wie in Abschnitt 4.1 diskutiert, sind Crosskonkordanzen intellektuell erstellte Verbindungen zwischen Termen zweier Thesauri oder Klassifikationen. Der häufigste Grund, warum auf intellektuell erstellte Crosskonkordanzen verzichtet wird, ist der nicht unerhebliche Aufwand. Schon deshalb muss eine andere Möglichkeit, den Transfer durchzuführen, zur Verfügung stehen: der Einsatz automatischer statistischer Verfahren. Dabei werden die jeweiligen Be-griffspaare nicht intellektuell festgelegt, sondern mithilfe mathematischer Verfahren berechnet. Die eingesetzten Algorithmen beruhen im Wesentlichen auf der Häufigkeit des „Zusammen-Vorkommens“ von Begriffen in sog. Parallelkorpora (Term-Kookkurrenzen), die als Trainingsdaten dienen. Parallelkorpora sind Dokumentsammlungen, deren einzelne Dokumente nach zwei Begriffssystemen indexiert sind (z. B. SWD und Thesaurus Sozialwissenschaften des IZ). Sie werden nicht durch intellektuelle Doppelindexierung neu geschaffen - was wiederum kostenintensiv wäre -, sondern man findet sie bereits vor und nutzt diesen Umstand. Die Ausgangssituation ist bei der Zusammenführung von Bibliotheksbeständen mit Fachdatenbanken in der Regel besonders günstig, da Informationszentren neben den Aufsätzen immer auch die selbstständige Literatur aufnehmen, die damit zumindest doppelt verschlagwortet vorliegt. So konnte für die Verbindung von SWD und dem Thesaurus Sozialwissenschaften des IZ in ViBSoz ein Parallelkorpus von 33.328 Dokumenten zugrunde gelegt werden. Aber auch die einzelnen Bibliotheksverbünde in Deutschland produzieren doppelte Indexierungen, die sich als Grundlage nutzen lassen.

Eine spezielle Art von Parallelkorpora lässt sich durch eine zusätzliche Freitextindexierung von Dokumenten erreichen, die bereits intellektuell mit einem Thesaurus/einer Klassifikation erschlossen wurden. Die errechnete Beziehung betrifft dann die Relation zum Freitextterm.

Auch der gleiche Text in mehreren Sprachen kann die Basis für einen Parallelcorpus bilden, in dem verschiedensprachliche Freitextterme aufeinander bezogen werden. Für das multilinguale IR ergänzt somit der gleiche bzw. hilfsweise ein ähnlicher Text in einer Fremdsprache den Parallelcorpus. Am

generellen Prinzip ändert sich dadurch nichts. So basieren alle bisher in ViB-Soz und ELVIRA getesteten Module auf einem Ansatz, der von Sheridan/Ballerini 1996 für das multilinguale Retrieval vorgeschlagen wurde.

Vereinfacht ausgedrückt gilt: Je häufiger ein Begriffspaar aus zwei Erschließungssprachen in einem Parallelcorpus vorkommt oder je häufiger sich Begriffe aus zwei unterschiedlichen Dokumenten aufeinander beziehen lassen (Term-Kookkurrenzen), desto wahrscheinlicher ist es, dass es sich um eine sinnvolle Verbindung handelt. Hinzu kommen Parameter wie die Größe des Korpus oder die Verteilung der Begriffe innerhalb des Textes. Eine ausführliche Beschreibung verschiedener Transferverfahren und ihrer Einsatzmöglichkeiten findet sich in Hellweg et al. 2001.

Für die Realisierung statistischer Transfermodule steht das Werkzeug „Jester“ (**J**ava **E**nvironment for **S**tatistical **T**ransf**ER**s) zur Verfügung, das im Kontext des Projekts ELVIRA erstellt und in CARMEN modifiziert wurde. JESTER erzeugt eine nach Wahrscheinlichkeit gewichtete Term-Term-Matrix. Zugelassen sind 1:1 und 1:m Relationen. Der Benutzer kann Schwellenwerte der Gewichtung angeben und bestimmte Terme (z. B. Allgemeinbegriffe) aus der Matrix ausschließen (siehe Hellweg et al. 2001).

Im Unterschied zu den Crosskonkordanzen in Abschnitt 4.1 ergeben sich die statistischen Transformationen nicht durch allgemeine intellektuell ermittelte semantische Beziehungen, sondern die Term-Term-Matrix spiegelt die errechneten Term-Kookkurenzen wider. Teilweise entsprechen sie den semantischen Beziehungen, die auch intellektuell ermittelt wurden. Es ergeben sich aber auch davon abweichende Relationstypen, die zu anderen relevanten Treffern führen.

Abfallpolitik	Abfallwirtschaft, Umweltpolitik	0.60.6
Jugendarbeitslosigkeit	Jugendli- cher,Arbeitslosigkeit	0.880.8 3
Wissensbasiertes System	Informationssys- temKünstliche Intelligenz	1.0 1.0

Abb. 4: Beispiele für statistische Transferbeziehungen, die den intellektuell ermittelten entsprechen (SWD ↔ Thesaurus Sozialwissenschaften des IZ)

Das folgende Beispiel stammt aus CARMEN, wo sowohl intellektuelle Crosskonkordanzen zwischen MACS (Mathematische Klassifikation) und

PACS (Physik) von Bibliothekaren erarbeitet als auch statistische Term-Term-Matrizen auf der Basis eines Parallelkorpus errechnet wurden (siehe Binder et al. 2002: Abschnitt 5.3.3 mit weiteren Beispielen und genauen Angaben zum Testdesign). Einzelne mathematische Methoden traten gehäuft in Verbindung mit inhaltlichen Forschungsgebieten auf, eine für das Retrieval wertvolle Relation, die bei der intellektuell erstellten Crosskonkordanz nicht berücksichtigt wurde.

PACS 62.30.+d	Mechanical and elastic waves; vibrations (Mechanische und elastische Wellen, Schwingungslehre)
MSC 74S15	Boundary element methods (Randelementmethode)

Abb. 5: Statistische Methodenbeziehungen MSC ↔ PACS

Aber auch statistische Kookkurrenzen, die sich nicht mehr systematisch erklären lassen, können zu relevanten Dokumenten führen, wie das folgende Beispiel aus einem Retrievaltest von sozialwissenschaftlichen Texten (Freitext) zeigt. Der Ausgangsbegriff ohne die Zuschaltung der Transferbegriffe war *Dominanz* und führte zu 16 relevanten Treffern.

Transferbegriffe	Dominanz, Messen, **Mongolei**, Nichtregierungsorganisation, Flugzeug, Datenaustausch, Kommunikationsraum, Kommunikationstechnologie, Medienpädagogik, **Wüste**
Zahl zusätzliche relevante Treffer	7
Anteil der zusätzlichen relevanten Treffer an den zusätzlichen Treffern	50 %
Erklärungstext	Mitglieder des Vereins wom@n reisten zur UNO-Frauenkonferenz nach Beijing. Auf der Fahrt durch die **Mongolei** und die **Wüste** ...

Abb. 6: Beispiel *Dominanz*

4.3 Statistisch ermittelte Crosskonkordanzen

Die in Abschnitt 4.2 als Alternative zum Einsatz intellektuell erstellter Crosskonkordanzen eingeführte statistische Transferbeziehung zeichnet sich nicht zuletzt durch die gefundenen Relationen aus, die nicht zwingend auf semantisch-pragmatischen Beziehungen basieren. Der Bearbeiter einer intellektuell erstellten Crosskonkordanz würde sie deshalb nicht angeben. *Wüste* und

Mongolei in Abb. 6 machen für den Benutzer, der die Termerweiterungen vor ihrer Verwendung prüfen möchte, in der Regel keinen Sinn, können aber dennoch zu relevanten Dokumenten führen. Deshalb liegt es nahe, Prozesse dieser Art als Hintergrundprozesse ablaufen zu lassen, vor allem dann, wenn zusätzlich zum statistischen Transfer eine intellektuell erstellte Crosskonkordanz zur Verfügung steht. Ein statistischer Transfer als Hintergrundprozess schließt die Möglichkeit nicht aus, dass der Benutzer über die Termerweiterungen informiert wird (wie bei ViBSoz) und die vom System vorgeschlagene Liste verändern kann.

Eine andere Ausgangssituation entsteht, wenn Kostengründe die Entwicklung einer intellektuell erstellten Crosskonkordanz verhindern. Die Term-Term-Matrix des statistischen Verfahrens kann dann - in der Regel nach einer zwischengeschalteten intellektuellen Überarbeitung - diejenigen semantisch-pragmatischen Relationen liefern, die Benutzer in jedem Thesaurus bzw. jeder Klassifikation erwarten.

In diesem Fall sprechen wir von statistisch erstellten Crosskonkordanzen. Sie wurden in CARMEN zusätzlich zur intellektuell erstellten Crosskonkordanz zwischen der mathematischen Klassifikation MSC und der physikalischen PACS erarbeitet. Das Motiv war hier die geringe Ausgefächertheit der Begriffe, die Klassifikationen inhärent ist.

In ELVIRA entstand aus Kostengründen eine statistische Crosskonkordanz für die Terminologie des VDMA. Sie ersetzt eine intellektuell erstellte Crosskonkordanz.

4.4 Zusammenspiel der Einzelverfahren

Die in den IZ-Projekten eingesetzten Heterogenitätsverfahren werden heute noch alternativ eingesetzt, d. h. eine Transformation besteht aus einem von möglicherweise mehreren für einen Übergang zur Verfügung stehenden Verfahren. Welche Transfers z. B. in ViBSoz zugeschaltet sind, entscheidet der Benutzer. Nur bei Nullantworten werden bei CARMEN systemseitig alle verfügbaren Transfers - in der Reihenfolge ihrer Relevanz - angestoßen, bis sich Dokumente nachweisen lassen.

Darüber hinaus wurden einige Ad hoc-Festlegungen zur Parametrisierung beider Einzelverfahren getroffen, ohne dass hierzu bereits systematische Untersuchungen vorliegen. So ist beim Einsatz intellektueller Crosskonkordanzen zu entscheiden, ob bzw. welche der Relationen Synonymie, Ober- / Unterbegriff, Ähnlichkeit standardmäßig zugeschaltet werden, um das bestmög-

liche Transferergebnis zu erreichen. Bei den statistischen Verfahren spielen die Schwellenwerte der Wahrscheinlichkeiten, die Verbindung mit linguistischen Verfahren (als vorgeschaltete Termbereinigung) und die Gewichtung der Art der Inhaltserschließung (z. B. Schlagworte höher gewichten als Stichworte) eine Rolle.

Auf der Basis der empirisch motivierten Optimierung der Einzelverfahren, die sich je nach Anwendungskontext und Fachgebiet deutlich unterscheiden wird, kann dann der eigentliche „Mehrwert“ der hier vorgeschlagenen Heterogenitätsbehandlung auf der Basis bilateraler Transfers zwischen einzelnen Dokumentenmengen modelliert werden. Ihr Vorteil liegt gerade darin, dass die unterschiedlichsten Mischformen möglich sind. So könnten die Transfermodule zwischen einem Dokumentbestand, der mit einer generellen bibliothekarischen Schlagwortliste wie SWD indexiert wurde und einem zweiten, dessen Indexierung auf einem speziellen fachspezifischen Thesaurus beruht, durch qualitative Verfahren wie Crosskonkordanzen aufeinanderbezogen werden. Die Einbindung der dritten Datenquelle (z. B. WWW-Angebote wie Clearinghouses oder graue Literatur) könnte wiederum über statistische Transfers (Term-Kookkurrenz) oder neuronale Netzwerke erfolgen. Die Verbindung zur Benutzerterminologie ließe sich abschließend mit einem probabilistischen Verfahren herstellen. Diese Möglichkeit, die einzelnen Stufen des Transfers jeweils differenziert an die speziellen Gegebenheiten anzupassen, ist ein wesentlicher Unterschied zu den bisherigen IR-Lösungen. Im gesamten Forschungskontext ist jedoch bisher völlig ungeklärt, wie die verschiedenen Verfahren kombiniert werden sollen, um größtmögliche Effizienz zu erreichen. Dass eine solche Kombination deutliche Vorteile verspricht, darauf verweisen die Ergebnisse der jährlich stattfindenden TREC- und DELOS–Evaluationen. Hier stellt sich immer wieder heraus, dass sich verschiedene IR-Verfahren nicht gegeneinander ausschließen, sondern unterschiedliche Mengen relevanter Dokumente liefern. Auch die ersten empirischen Ergebnisse aus CARMEN mit dem Vergleich alternativer Zugänge weisen in diese Richtung (siehe Binder et al. 2002).

Bereits heute lässt sich feststellen, dass im ViBSoz- und CARMEN-Kontext empirische Belege für die Qualitätsverbesserung der Recherche durch den Einsatz der Transfermodule in ihrer jetzigen - gegenüber der Modellvorstellung noch sehr beschränkten - Architektur nachweisbar sind. Der eigentliche Mehrwert des Verfahrens wird sich jedoch erst nach der empirischen Fundierung verschiedener Parametrisierungen der Einzelmodule und der möglichen Kombinatorik der unterschiedlichen Transfertypen entfalten können. Hierzu ist am IZ eine Dissertation in Arbeit, deren Ergebnisse im Frühjahr 2004 vorliegen dürften.

5 Transfer-Architektur und Einbettung in dezentrale und zentralistische Ansätze

Varianten der beschriebenen Transfermodule wurden bisher in vier verschiedene Architekturen eingebettet (ELVIRA, CARMEN, ViBSoz, ETB). Zwei davon sind heute im praktischen Einsatz (CARMEN HYREX und die ETB-Transfers kamen nicht über die Prototypphase hinaus).

Da die Transfermodule unabhängig von der Systemarchitektur des Gesamtsystems sein sollten, um portabel zu sein, war eine möglichst unabhängige Integrationsebene zu finden. Sie setzt als Termerweiterungs- bzw. Veränderungsverfahren bei den Eingabebegriffen des Nutzers an und verteilt die Ergebnisse auf Untermengen der Datenbestände. Diese zwei Merkmale genügen, um die Grundidee der vorgeschlagenen Heterogenitätsbehandlung in den unterschiedlichsten Architekturmodellen verwenden zu können.

Abb. 7 zeigt exemplarisch die Einbindung in die CARMEN-Architektur, die wie ViBSoz dezentrale Datenbanken bzw. Web-Angebote miteinander verbindet: Ein Benutzer der Universität Köln ist z. B. mit den dortigen OPACs vertraut und formuliert deshalb seine Anfrage mit den Begriffen des SWD, will aber gleichzeitig in der SOLIS-Datenbank suchen und in einem dritten Dokumentenbestand mit wiederum unterschiedlicher Inhaltserschließung. Die Anfrage wird unverändert an die Universität Köln geschickt (Recherche Bestand B), als Variante 2 nach einem bilateralen Transfer SWD ↔ Thesaurus Sozialwissenschaften des IZ (z. B. intellektuell erstellte Crosskonkordanz) an SOLIS (A) gerichtet usw.

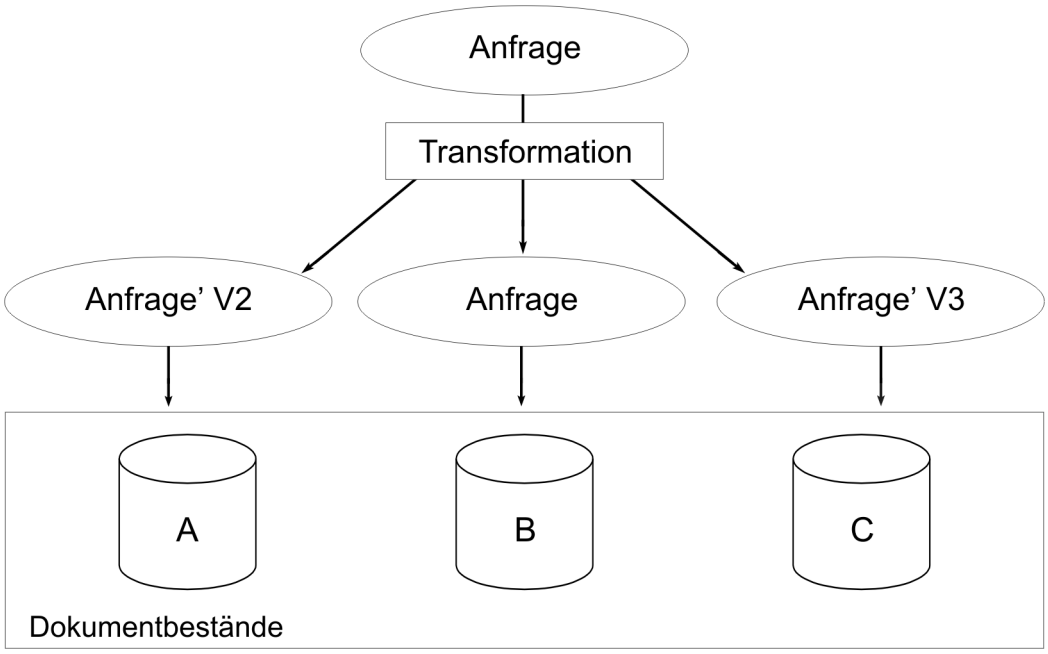

Abb. 7: Anfrage – Transformation aus Hellweg et al. 2001

Beim EU-Projekt ETB ging es im Unterschied zu der dezentralen Brokerarchitektur von CARMEN und ViBSoz um eine zentralistische Architektur. Die einzelnen Anbieter von Bildungsdatenbanken (Site A, B usw.) wollten ein zentrales Angebot aufbauen mit einer einheitlichen Zugangssprache (ETB Native Repository). Eigentlich widerspricht dieser Ansatz den hier und in Krause/Niggemann/Schwänzl 2003 vorgetragenen Überlegungen, da er vordergründig nur auf eine neue übergeordnete Standardisierung setzt. Heterogenitätsmodule kamen dennoch für alle Altbestände, die die neue ETB-Erschließungssprache als Zielpunkt des Transfers auf Dokumentenebene einsetzten zum Zuge und für all jene, die zweigleisig bleiben wollten, d. h. ihre eigene Form der Inhaltserschließung beibehielten und die Arbeit für eine Doppelindexierung scheuten. Auch in diesem Fall erzeugen die Transferkomponenten die Zielstrukturen.

Das Beispiel zeigt, dass auch bei zentralistischen Architekturen, die an einer zentralen Datenbank und einer einheitlichen Beschreibungssprache festhalten, Transfermodule ihren Platz haben.

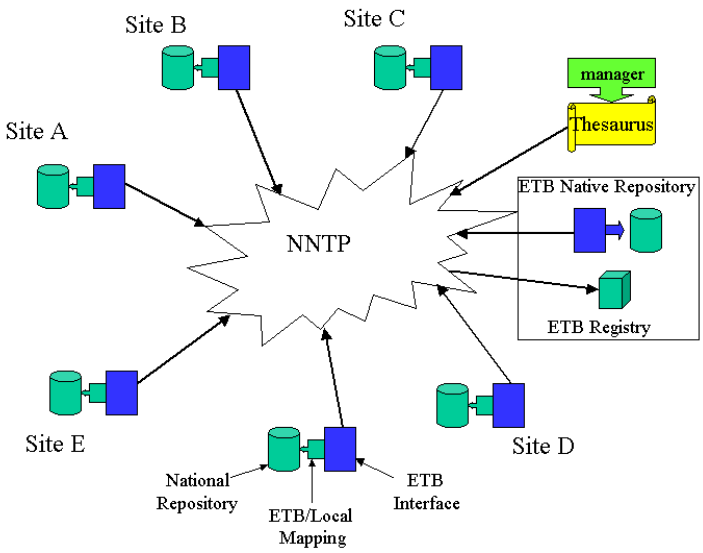

Abb. 8: ETB-Architektur aus Bandholm/Lund p. 7

6 Wie konzeptualisiert der Benutzer seine Anfrage an einen dezentralen, heterogenen Informationsraum?

Solange der Benutzer einem Dokumentenbestand gegenübersteht, den er als einen Informationstyp empfindet, wird er die Wünsche an die Benutzungsoberfläche aus der Vorstellung eines exemplarischen Vertreters des Dokumentenbestandes ableiten. Bei der Literaturrecherche erwartet er bestimmte Strukturen und Merkmale (Autor, Titel, Schlagworte usw.), die über alle Dokumentenbestände verschiedener OPACs oder Fachinformationssysteme hinweg angenommen werden. Die einheitliche Rechercheoberfläche muss dieser Erwartungshaltung Rechnung tragen.

Schon bei der Ansetzung verschiedener Inhaltserschließungen in einer digitalen Bibliothek, derer sich der Benutzer bewusst ist, stellt sich aber die Frage nach dem anzubietenden Einstieg. ViBSoz geht z. B. davon aus, dass Benutzer, die an der Universität Köln suchen und mit den dortigen OPACs vertraut sind, im SWD-Thesaurus ihre Suchbegriffe auswählen (explizit über das Thesauruswerkzeug oder implizit). Die Transformationen richten sich dann auf die dem Benutzer nicht vertrauten Inhaltserschließungssysteme. SWD prägt somit das exemplarische Vorbild des Benutzers, das durch den Anfragebildschirm zu unterstützen ist. Benutzer von SOLIS legen dagegen den Thesaurus Sozialwissenschaften des IZ als ihren Standardeinstieg zugrunde.

Man könnte zusätzlich zu diesen Einstiegen davon ausgehen, dass es auch Benutzer gibt, die ihr Informationsbedürfnis abstrahiert von den verschiedenen Inhaltserschließungssystemen des dezentralen Informationsraums konzeptualisieren wollen, auch wenn sie zumindest eines der Inhaltserschließungssysteme genauer kennen. Sie würden dann quasi ihr eigenes Begriffssystem als Startpunkt aller Transformationsprozesse, die systemseitig angestoßen werden, setzen. Dies klingt plausibel und scheint problemlos umsetzbar. Schwieriger wird es jedoch, wenn neben textuelle Dokumente Faktendaten treten, wie dies bei ELVIRA der Fall ist.

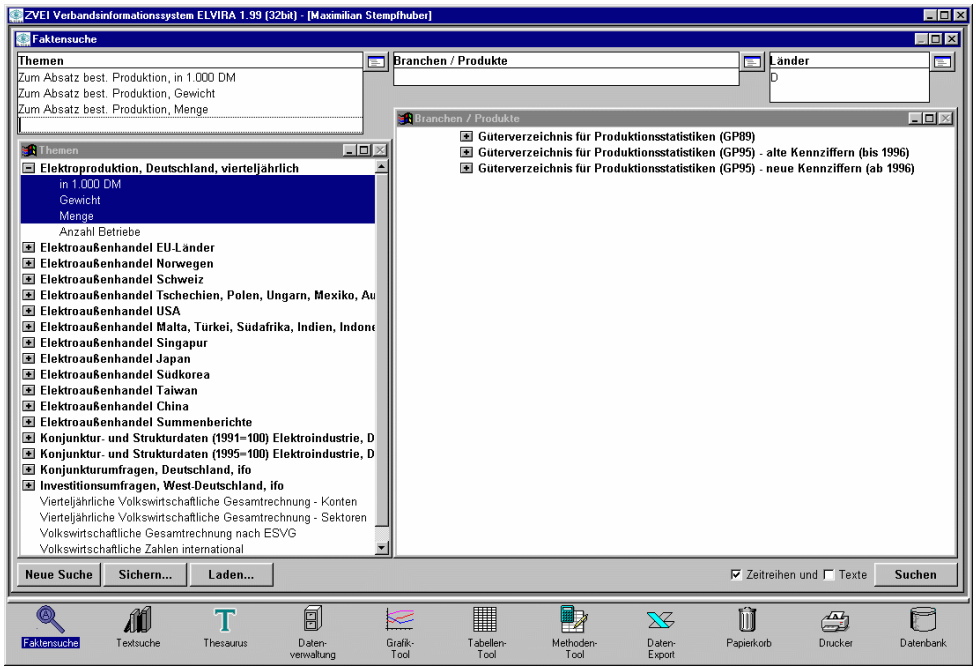

Abb. 9: ELVIRA-Faktenrecherche, Stempfhuber 2003

ELVIRA wurde in der ersten Stufe als Fakteninformationssystem für die Firmen der Elektronikindustrie entwickelt. Es sollte einen benutzerfreundlichen Zugang zu Produktions-, Außenhandels-, Konjunktur- und Strukturdaten ermöglichen. Der Recherchezugang ist durch die spezifische Datenstruktur und die Art der Inhaltserschließung der bei den Verbänden intellektuell ermittelten Daten gekennzeichnet. Die Zeitreihen-Faktentabellen werden nicht über ihre Zellenwerte und den Tabellennamen angesprochen, sondern indirekt über intellektuell vergebene Deskriptoren zu den drei Kategorien: betroffenes Thema (z. B. Export), Branche/Produkt (z. B. Mikrowellengeräte) und Land (s. Scheinost et al. 1998, Stempfhuber 2003).

Benutzertests zeigten trotz hoher Akzeptanz rasch, dass die Verbandskunden neben den Zeitreihen andere Informationsquellen zur Lösung ihrer Problemstellungen fordern. Beim VDMA waren dies z. B. textuelle Außenwirtschafts-

informationen zur qualitativen Marktbeurteilung. Deshalb wurde ELVIRA um einen Textzugang erweitert und eine integrierte Suche auf der Basis des in den vorangegangenen Abschnitten vorgestellten Modells von Transfermodulen zugelassen.

Einen Sinn macht die Integration von Fakten und Texten allerdings erst, wenn eine zweite Lücke der heute üblichen Zugangssysteme geschlossen wird. Es muss ein adäquates Modell für die Behandlung des Problems der Rechercheformulierung durch den Benutzer gefunden werden – oder einfacher ausgedrückt: Wie sieht der einheitliche Anfragebildschirm in diesem Fall aus?

Zusammen mit Nutzern und Informationsvermittlern bei den Verbänden wurden zunächst die relevanten Textbestände durch die Auswertung von Anfragen an die Verbände und durch Benutzerinterviews ermittelt. Die Untersuchung zeigte, dass es sowohl Fragestellungen gab, bei denen der Benutzer entweder *nur* nach Zeitreihen oder *nur* nach Texten sucht, dass aber auch Mischformen eine Rolle spielen. Die beiden Reinformen, die Fakten- und die Text-Recherche, dienten als Ausgangspunkt für die Integration von Text- und Faktenrecherche.

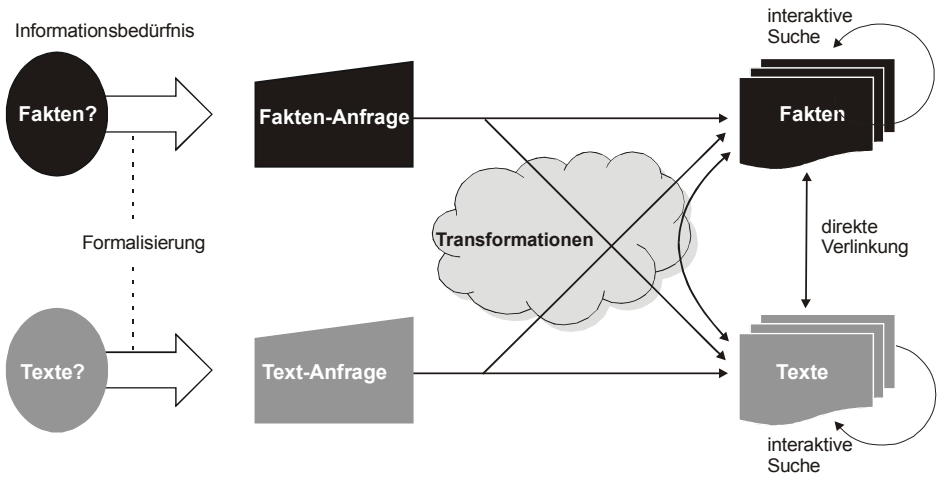

Abb. 10: Reinformen

Für unser Problem sind die Fälle von besonderem Interesse, bei denen der Benutzer ausgehend von einem Ergebnistyp (z. B. Texten) plötzlich auch Ergebnisse anderen Typs (z. B. Zeitreihen) möchte. Dabei macht die Gestaltung des Bildschirms der Ausgangsfrage bei den folgenden Übergangstypen keine Probleme:

- Fakten-Anfrage → (Fakten- und) Text-Ergebnis
- Text-Anfrage → (Text- und) Fakten-Ergebnis
- Fakten-Ergebnis → Text-Ergebnis
- Text-Ergebnis → Fakten-Ergebnis

All dies sind keine „echten“ Integrationen auf der Ebene des einheitlichen Anfragebildschirms. Eine echte Integration liegt dann vor, wenn der Benutzer sein Informationsbedürfnis völlig abstrakt formuliert (s. Abb. 11).

Die ersten empirischen Untersuchungen ergaben Hinweise, dass solche Bedürfnisse im Anwendungsfall von ELVIRA vorkommen (siehe Krause/Mandl/Stempfhuber 1997). Der Versuch, diese echte Integration zusammen mit den Benutzern als Anfragebildschirm zu erarbeiten, zeigte aber, dass im konkreten Fall einer definierten Recherche der Benutzer immer entweder vom Text oder von der Faktendarstellung als Vorstellung ausgeht. So gibt er zwar an, er wolle „alles“ zu einem Land wissen. Bei der Rückfrage, wie er sich dazu den idealen Anfragebildschirm wünscht, wechselt er jedoch wieder zu einer konkreten Ausgangsfrage im Text- oder Faktenmodus. Nach den bisherigen Erfahrungen spricht deshalb vieles dafür, dass bei der Suchformulierung der Text-Fakten-Integration bereits in Beispielen gedacht wird und damit das abstrakte Informationsbedürfnis auf der Ebene der Suchformulierung keine Entsprechung mehr hat.

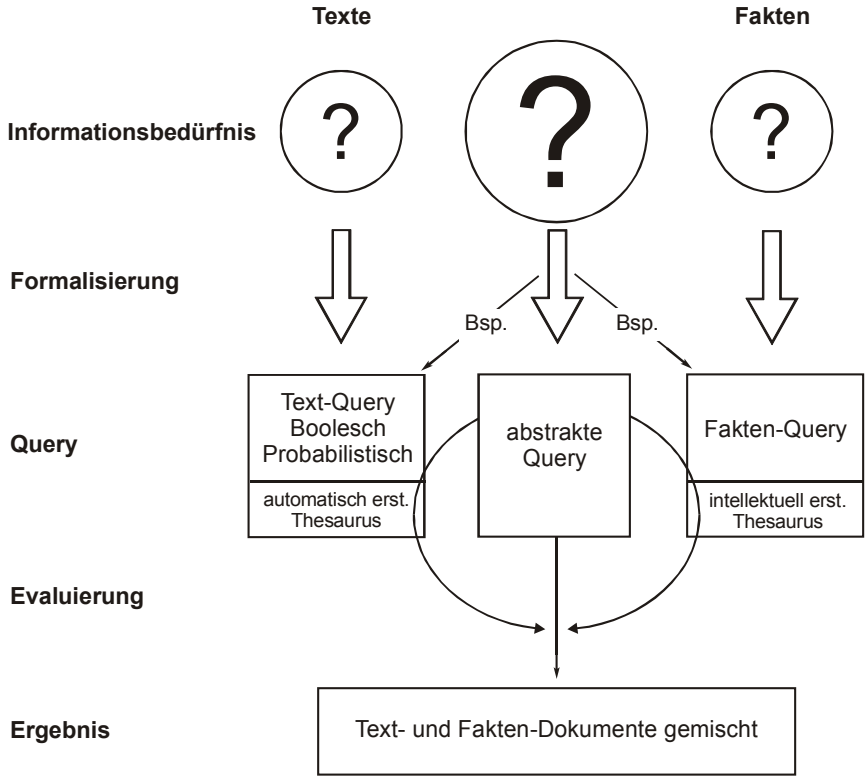

Abb. 11: „Echte“ Integration

Damit würden die jeweiligen Reinformen der Recherche als Anfragebildschirme genügen. Ob sich diese Beobachtung in anderen Heterogenitätskontexten bestätigt, muss abgewartet werden.

7 Transfermodule als erster Baustein eines Schalenmodells der Informationsverarbeitung

Bilaterale Transfermodule sind von der Modellbildung her sehr einfache Grundbausteine, die jedoch durch ihre Variationsbreite und die kaskadierende Anwendung der Einzelelemente schnell sehr komplex werden können. Sie sind bei den bisherigen Anwendungen mit ihrer beschränkten Vielfalt der zu integrierenden Inhaltserschließungstypen noch übersichtlich analysierbar und handhabbar. Bei der sehr viel größeren Anzahl von Variationen, die uns erwarten, wenn wir die Integrationsmöglichkeiten des WWW ernst nehmen, dürfte sich das jedoch rasch ändern.

Deshalb braucht das vorgeschlagene Modell abstraktere Ordnungsansätze, die auf einem höheren Niveau der Zusammenfassung arbeiten.

Dies soll das Schalenmodell leisten. Es wurde zusammen mit einem noch nicht weiter ausgearbeiteten Modell der Transfermodule erstmals für die Daten des IZ Sozialwissenschaften vorgeschlagen und bezieht neben der informationswissenschaftlichen Ebene organisatorische und wissenschaftspolitische Dimensionen mit ein (Krause 1996). Es ergänzt die bilateralen Transfermodule um einige zusätzliche Annahmen: Verschiedene Niveaus der Inhaltserschließung und Dokumentenrelevanz werden zu sog. Schalen zusammengefasst, die untereinander durch höherstufige Transfermodule verbunden werden (genauer Krause 1999). Alle früheren Veröffentlichungen zum Schalenmodell ließen allerdings die in den vorangegangenen Abschnitten skizzierte Problematik der gegenüber dem traditionellen IR veränderten Behandlung der Vagheit zwischen den einzelnen Schalen noch unspezifiziert und sind somit um die obigen Ausführungen zur bilateralen Heterogenitätsbehandlung zu ergänzen.

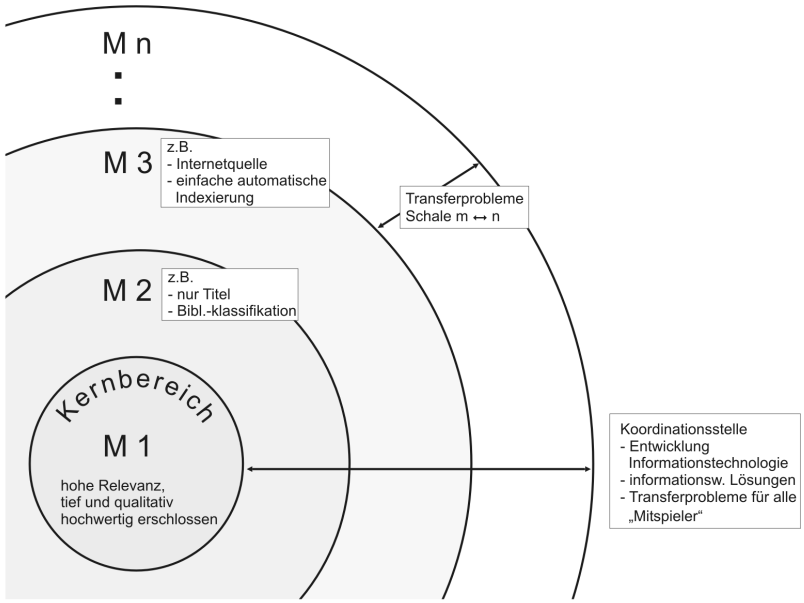

Abb. 12: Schalenmodell aus Krause 1999

Wie viele Schalen angesetzt werden und welche Merkmale sie definieren, richtet sich nach den Gegebenheiten eines Fachgebiets und den sich beteiligenden Anbietern. Es ist zudem damit zu rechnen, dass innerhalb einer Schale Transformationen zwischen Teilbeständen angesetzt werden müssen, wenn partiell das für das Schalenmodell konstitutive Zusammenwirken von Relevanz und Erschließungstiefe nicht vorliegt. Damit lässt sich das Schalenmodell - und somit die Vorteile seiner weitergehenden Annahmen - auch dann anwenden, wenn die Kriterien Relevanz und Qualität der Inhaltserschließung sich nur teilweise parallelisieren lassen. Dies dürfte eher die Regel als die Ausnahme sein.

Generell wissen wir noch recht wenig über den richtigen Aufbau eines Schalenmodells. Sicher ist nur, dass sich die Lösungen in verschiedenen Anwendungsfeldern und abhängig von den zu integrierenden Dokumententypen deutlich unterscheiden werden. Deshalb müssen die möglichen Transferkomponenten systematisch an verschiedenen Textmaterialien analysiert und exemplarisch in prototypische Lösungen umgesetzt werden, bevor eine weitere Spezifizierung möglich ist.

Das Konzept der bilateralen Transfermodule ist heute - von der Modellbildung und von den praktischen Einsatzmöglichkeiten her - so weit fortgeschritten, dass es sich konkret bei digitalen Fachbibliotheken, Fachportalen und Informationsverbünden mit Gewinn einsetzen lässt. Die weitergehenden Anforderungen des Schalenmodells geben demgegenüber heute noch eher eine Denkrichtung an, die den weiteren Ausbau eines Richtungswechsels in der Fachinformation kennzeichnen soll, den ich zum Abschluss noch einmal wiederholen möchte:

„Standardisierung ist von der verbleibenden Heterogenität her zu denken“.

8 Literatur

Bandholm, Anders; Lund, Tommy Byskov (2000): WP9 Architecture of the Metadata Networking Infrastructure, European Treasury Browser, Deliverable No. D9.1.
URL: www.eun.org/etb/Output and Documents

Belkin, Nicholas J. (1996): Intelligent Information Retrieval: Whose Intelligence? In: Krause, Jürgen; Herfurth, Matthias; Marx, Jutta (Hrsg.): Herausforderungen an die Informationswissenschaft. Konstanz 1996, S. 25 - 31.

Binder, Gisbert; Marx, Jutta; Mutschke, Peter; Riege, Udo; Strötgen, Robert; Kokkelink, Stefan; Plümer, Judith (2002): Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht; Nr. 24)
URL: http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_24.pdf

DIN 820-1; Ausgabe 1994-01: Normungsarbeit: Grundsätze. Berlin: Beuth.

DIN-SICT (2002) „Strategie für die Standardisierung der Informations- und Kommunikationstechnik (ICT)“. Berlin: DIN.

Hellweg, Heiko; Krause, Jürgen; Mandl, Thomas; Marx, Jutta; Müller, Matthias N.O.; Mutschke, Peter; Strötgen, Robert (2001): Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht; Nr. 23)
URL: http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_23.pdf

Ingwersen, Peter (1996): The Cognitive Framework for Information Retrieval: A Paradigmatic Perspective. In: Krause, Jürgen; Herfurth, Matthias; Marx, Jutta (Hrsg.) Herausforderungen an die Informationswissenschaft. Konstanz 1996, S. 25 - 31.

Kluck, Michael (2001): Report on results and usability of cross-concordances lists of controlled terms and authority lists, European Treasury Browser, Deliverable No. 6.1.
URL: www.eun.org/etb/Output and Documents

Kluck, Michael; Strötgen, Robert (2002): Report on the solution for transfer components and description of the underlying software, European Treasury Browser, Deliverable No. D6.2.
URL: www.eun.org/etb/Output and Documents

Krause, Jürgen (2000): Integration von Ansätzen neuronaler Netzwerke in die Systemarchitektur von ViBSoz und CARMEN. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht; Nr. 21).
URL: http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_21.pdf

Krause, Jürgen (1999): Sacherschließung in virtuellen Bibliotheken – Standardisierung von Heterogenität. In: Rützel-Banz, Margit (Hrsg.): Grenzenlos in die Zukunft. 89. Deutscher Bibliothekartag in Freiburg im Breisgau 1999. Frankfurt am Main: Klostermann, S. 202 - 212.

Krause, Jürgen (1996): Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung („Schalenmodell“). Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht; Nr. 6).
URL: http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab6.pdf

Krause, Jürgen; Elisabeth Niggemann; Roland Schwänzl (2003): Normierung und Standardisierung in sich verändernden Kontexten: Beispiel: Virtuelle Fachbibliotheken. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie 50, Nr. 1, S. 19-28.

Krause, Jürgen; Schwänzl, Roland; Plümer, Judith (2000): Content Analysis, Retrieval and Metadata: effective Networking for Mathematics, Physics and Social Sciences. In: Blasius, Jörg; Hox, Joop; Leeuw, Edith de; Schmidt, Peter (Hrsg.): Social Science Methodology in the New Millennium: Proceedings of the Fifth International Conference on Logic and Methodology, Cologne, October 3-6, 2000. CD-ROM. Amsterdam: TT-Publikaties.

Krause, Jürgen; Mandl, Thomas; Stempfhuber, Maximilian (1997): Text-Fakten-Integration in ELVIRA. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht; Nr. 12).
URL: http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab12.pdf

Mandl, Thomas (2001): Tolerantes Information Retrieval. Neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. Dissertation. Konstanz: UVK, Univ.-Verl. (Schriften zur Informationswissenschaft; Bd. 39).

Müller, Matthias N.O.; Marx, Jutta (2001): The Social Science Virtual Library Project: Dealing with Semantic Heterogeneity at the Query Processing Level. S. 19 - 24. In: Proceedings of the Third DELOS Network of Excellence Workshop „Interoperability and Mediation in Heterogeneous Digital Libraries”, Darmstadt, Germany; 8 - 9 Septem-

ber 2001. Sophia-Antipolis, USA: ERCIM. (DELOS Network of Excellence on Digital Libraries Workshop Series)

Mutschke, Peter (2000): Cognitive and Social Structures in Social Science Research Fields and their Use in Information Systems. - 5th International Conference on Logic and Methodology Social Science „Methodology in the New Millenium“, Köln, 03. - 06. Oktober 2000.

Mutschke, Peter; Quan-Haase, Anabel (2001): Collaboration and Cognitive Structures in Social Science Research Fields: Towards Socio-Cognitive Analysis in Information Systems. In: Scientometrics 52, Nr. 3, S. 487 – 502.

Ronthaler, Marc, Zillmann, Hartmut (1998): Literaturrecherche mit OSIRIS; ein Test der OSIRIS – Retrievalkomponente. In: Bibliotheksdienst, 32(7), S. 1203-1212.

Scheinost, Ulrich; Haas, Hansjörg; Krause, Jürgen; Lindlbauer, Jürg D. (1998) (Hrsg.): Marktanalyse und Marktprognose: Das ZVEI Verbandsinformationssystem ELVIRA. Bonn: Informationszentrum Sozialwissenschaften (IZ-Forschungsberichte; Bd. 2). Bonn.

Schöning-Walter; Christa (2003): Die DIGITALE BIBLIOTHEK als Leitidee: Entwicklungslinien in der Fachinformationspolitik in Deutschland. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie, 50, Nr. 1, S. 4 - 12.

Sheridan, Páraic; Ballerini, Jean Paul (1996): Experiments in Multilingual Information Retrieval using the SPIDER system. S. 58 - 65. In: Frei, Hans-Peter; Harman, Donna; Schäuble, Peter; Wilkinson, Ross (Hrsg.): SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 18 - 22, 1996, Zurich, Switzerland. New York: Association for Computing Machinery; Konstanz: Hartung-Gorre.

Stempfhuber, Maximilian (2003): Objektorientierte Dynamische Benutzungsoberflächen - ODIN: Behandlung semantischer und struktureller Heterogenität in Informationssystemen mit den Mitteln der Softwareergonomie. Bonn: IZ Sozialwissenschaften. (IZ-Forschungsbericht; Bd. 6).

Stempfhuber, Max; Hellweg, Heiko; Schaefer, André (2002): ELVIRA: User Friendly Retrieval of Heterogenous Data in Market Research. S. 299 - 304. In: Callaos, Nagib; Hernandez-Encinas, Luis; Yetim, Fahri (Hrsg.): SCI 2002: The 6th World Multiconference on Systemics, Cybernetics and Informatics; July 14 - 18, 2002, Orlando, USA; Proceedings, Vol. I: Information Systems Development I. Orlando.

Strötgen, Robert (2002): Meta-Data Extraction and Query Translation. Treatment of Semantic Heterogeneity. S. 362 - 373. In: Agosti, Maristella; Thanos, Constantino (Hrsg.): Research and Advanced Technology for Digital Libraries: 6th European Conference, ECDL 2002, Rome, Italy, September 16 - 18, 2002; Proceedings. Berlin: Springer.

Zillmann, Hartmut (1997): OSIRIS. Ein gemeinsames Projekt der Universitätsbibliothek Osnabrück und des Instituts für Semantische Informationsverarbeitung der Universität Osnabrück (ISIV).Osnabrück 1997.
URL: http://www.ub.uni-osnabrueck.de/osiris/osiris_papers.html